Question(s):			**Meeting, date:**	VQEG, Ghent 22-26 Sept 2008	
Study Group:		**Working Party:**		**Intended type of document** (R-C-TD):	C
Source:	Psytechnics Ltd				
Title:	Issues in HDTV test plan v.2.2 proposals				
Contact:	Quan Huynh-Thu Psytechnics Ltd UK		Tel:+44 1473 261 800 Fax: +44 1473 261 880 Email: quan.huynh-thu@psytechnics.com		
Contact:			Tel: Fax: Email:		
Please don't change the structure of this table, just insert the necessary information.					

Summary

A number of issues need to be addressed in the HDTV test plan before it can be implemented. The document VQEG_HDTV_testplan_v_2_2_proposals.doc provides the latest version of the test plan, together with different proposals that have been made by different parties.

This document:

- lists additional issues not previously identified
- discusses the proposals made in v.2.2
- asks for clarification of some of these proposals
- makes additional proposals

Issues to discuss based on HDTV test plan v.2.2 proposals

Text highlighted in yellow is taken as is from the test plan.

Text highlighted in green indicates a new proposal.

Section 1 - Introduction

"It is also proposed that a test of currently standardized standard definition models be tested for their extensibility to High Definition TV."

When the HDTV project started a few years ago, there was a need from the industry to check quickly if the existing ITU-T J.144 models could be validated for HDTV. Since this effort did not really kick off, it has now become obsolete. All J.144 models are now at least 5 years old. Technology and knowledge in the field have evolved and we think it would be a waste of resource to have an extra work item dedicated to the validation of these models (unless someone volunteers

to do that job…). Nothing prevents the authors of these models to re-submit them (as is or modified/improved) to the HDTV project as a new model.

We support suppressing that sentence.

Section 2.3 – Test Schedule

- Point 7 "Sample video sequences agree upon":
    - Agreed by (between) whom?
- In the current schedule, proponents do the following steps in order: (1) submit their model to ILG, and at the same time submit their subjective data and PVSs used in their subjective test to ILG and other proponents, (2) run their model on all PVSs used in all subjective tests, (3) submit their objective scores to ILG and other proponents.
    - Who will check that the objective scores submitted by proponents are correct? The ILG must do this check (e.g. random check by running the proponent's model on a subset of PVSs or full check by running the proponent's model on all PVSs, then making sure the objective scores match values previously submitted by proponent).
- If the test plan considers the use of a similar software approach than for MM tests for presentation of the videos in the HD tests, then finalization of the software needs to be included in the schedule. If different test labs are allowed to use different presentation systems, this milestone is not required in the schedule.

Section 3.1 – Model Type

"Proponents may submit one model of each type (FR, RR, NR) to apply to all video formats (1080i50, 1080i60, 1080p30, and 1080p25). Thus, any single proponent may submit up to a total of five different models."

The five models refer to: one full reference, one no reference, and one for each of the three reduced reference information bit rates given in the test plan (i.e. 56 kbit/sec, 128 kbit/sec, 256 kbit/sec).

According to this text, an FR model must be applicable to all four video formats (1080i@50Hz, 1080i@60Hz, 1080p25 and 1080p30).

We think this constraint is too restrictive:

- A model developer might be interested in applications using progressive video only so will not develop or tune the model for interlaced video. In that case, the current test plan would not allow such model to be submitted to the HD validation test.
- This restriction will likely limit the number of proponents and therefore the number of subjective databases

We propose that proponents can choose which format(s) their model will address.

Section 3.2 – Full Reference Model Input and Output Data Format

"Each line may also optionally contain calibration values."

What is the reason to allow manual input values for calibration/registration?

FR and RR models are supposed to handle the spatio-temporal misalignment between source and processed videos so they should be given only the names of the files as inputs.

We propose to remove calibration/registration values as inputs to the FR model. In any case, such calibration/registration values passed as inputs to the models can only address constant delay, shift, gain and offset. So they wouldn't be useful to models to address variable delay as considered in the HD test plan.

Section 4.3 – Test Design

It was agreed in Kyoto that each subjective test will use the identical number of 162 PVSs, i.e. 9 SRCs x 18 HRCs (including hidden reference).

In v.2.2, additional text proposes that if a lab has access to many more SRCs, then not all SRCs are processed through all the same HRCs. This proposal seems reasonable since the number of subjective tests is likely to be limited. However we believe that it is crucial that all subjective tests still use the identical number of PVSs (i.e. 162). This point is particularly crucial as the number of tests to validate the models is likely to be small. The example of a test design using 171 PVSs provided in the proposal text of Section 4.3 in HDTV test plan v.2.2 is therefore not acceptable. If a non-rectangular design is used (where not all the SRCs are processed through all the same HRCs) then the final number of PVSs (including the hidden reference of each SRC) must be 162.

We therefore do not support the alternative proposed text in v.2.2.

We propose that if not all SRCs are processed through all the HRCs then the final number of PVSs should still be 162. However, it must be noted that in this case, a per-HRC or per-SRC data analysis can not be applied because the average value (per HRC or per SRC) would be the average over a different number of files, depending on the HRC or SRC. The analysis per HRC or per SRC would therefore not be meaningful.

Section 4.4.3 – Display Specification and Set-up

The opening paragraph of Section 4 specifies that subjective tests will deploy a variety of display technologies. Furthermore, Section 4.4.2 states that each test lab can use its own video display.

We believe that it is possible to run subjective tests with high-end consumer-grade monitors when the video material is in progressive format. From the list of TCO-certified monitors sent by ACREO (during the MM project), we have identified one monitor that is suitable for running HD tests: BenQ FP241WZ (or FP241W).

We propose that test labs can use a BenQ FP241WZ or FP241W LCD monitor for subjective tests using progressive video material (i.e. 1080p@25 or 1080p@30 test):

The advantages of such monitor are:

- Affordable price (about UK£500 so should be even cheaper in some other geographical areas)
- TCO 06 certified
- Response time: 6ms GTG
- 1920x1200 native resolution
- 1:1 display mode (1 image pixel = 1 display pixel) that won't apply any scaling on the picture
- Allows disabling post-processing
- Monitor won't add any additional post-processing if video is progressive
- Representative technology used in real-world applications

Section 4.4.3 – Display Specification and Set-up

Concerning the possibility to use a progressive display to present interlaced video material, the following proposal was made in v.2.2: "If the native display of the monitor is progressive and thus performs de-interlacing, then if 1080i SRC are used, the test video sequences must be de-interlaced before it is sent to the monitor. This de-interlaced video files must be made available (i.e., to proponents and ILG). The interlaced files will be used by the model. The de-interlaced files are to be made available for later studies and analysis of the influence of the de-interlacing on perceived quality. These studies constitute supplementary analysis resulting from the HDTV testing, intended to guide future testing."

- We ask clarification on this proposal (and especially on the sentence "The interlaced files will be used by the model ") as we do not understand the logic of the proposal: if the interlaced videos are de-interlaced before being input to the monitor (and therefore viewers will see progressive video), then why should models receive interlaced videos as inputs?
- This proposed text contradicts the following statement in Section 7.1.2 "Where a progressive display is used and the test sample requires de-interlacing, then this de-interlacing will be performed offline, and the model will be given the same de-interlaced sample as is shown to the viewer"
- The benefit of an off-line de-interlacing of the interlaced videos prior to sending them to the progressive display is to "control" the de-interlacing, since a defined de-interlacing process is applied (rather than leaving the monitor to apply its proprietary and unknown de-interlacing). However, models should then also be provided the de-interlaced videos as inputs.
- We propose the alternative easier solution that a CRT monitor should be used for 1080i tests and a progressive (e.g. LCD) monitor should be used for 1080p tests.

Section 7.2 - HRC Constraints and Sequence Processing

From reading the sentence "These error conditions must include the following", one can understand that any HRC must include coding AND pre/post-processing AND transmission errors.

We think that this was not the initial intent. From the discussions in Kyoto, we think that the intention of the sentence was to make sure that no other (unknown) type of errors (other than the ones listed) could be used in a HRC.

We propose to replace the sentence by "These error conditions are limited to the following"

Section 7.2.4 - Frame Freezing and Frame Skipping

The current text (as well as the proposed replacement text) allows re-buffering (as it allows freezing alone).

Is this really the intent?

We can not recall that re-buffering was ever discussed as an HRC in HDTV test plan.

Section 7.2.4 - Frame Freezing and Frame Skipping

The current replacement text reads "Frame freezing and frame skipping events are constrained primarily by the subjective testing methodology agreed upon herein. Because the SRC and PVS must have the same length (10 seconds), some extra content or missing content may result at the end of the video sequence. The maximum length of a frame freezing or frame skipping event is naturally limited by this length constraint on the PVS."

We see a problem with the last sentence. A specific limitation on the length of freezing and skipping is still required. Without a specific limitation such as the one in the current text ("The maximum of total freeze is 25% of the total length of the sequence"), the replacement text would allow a case where there is a freezing of 8 seconds followed by a skipping of 8 seconds. In the end, both the SRC and PVS will be 10-second long and equal in length and can even match on the last frame.

We propose that a specific limit be clearly mentioned.

We propose that freezing with skipping vs. freezing without skipping (re-buffering) be clarified.

Section 7.2.5 - Rewinding

We support leaving the text as is in v.2.1: rewinding is not allowed in HD tests.

Section 7.2.6 - Frame rates

We agree with the comment in v.2.2. The text in this section does not integrate the decisions made at the Kyoto meeting.

From the decisions made during the Kyoto meeting:

- 1080p: only 25 or 30 fps are used.
- 1080i: 50Hz or 60Hz.

Section 7.3.1 – Pre-processing

3:2 pull-down is mostly used to convert a 24 fps movie into a 30 fps video. Therefore this processing would be used to produce a SRC not as part of a HRC. It cannot be used to produce a PVS since the SRC and PVS must have the same frame rate.

We propose to remove 3:2 pull down from the list of pre-processing steps in producing the HRCs.

Section 7.3.3 – Distribution of HRCs

About the proposal to have at least 3 HRCs containing each codec:

- See previous comment on Section 7.2.1

About the proposal on transmission errors:

- It would be desirable to have transmission errors included in this round of testing
- On the other hand, model validation will require a certain amount of data with this type of error to have a meaningful validation. Producing videos at1080 (i or p) resolution with transmission errors is problematic as it requires capture of video in uncompressed format at 1080 (i or p) resolution at 25 or 30fps. This requires very expensive equipment (such as DVS recorder) to produce such material. Since most of material will be produced by proponents, it must be checked whether most proponents can produce this type of error condition.

Section 8.1 – Calibration Constraints

About the temporal registration:

- The initial text in version 2.1 of the HDTV test plan was very similar to the MM test plan in order to consider variable delay (caused by frame freezing/skipping).
- The major difference between HD and MM is that in HD the frame rate remains constant and at the maximum value (i.e. there is no HDTV at 15 frames per second)
- However, events of frame freezing and frame skipping (allowed in the test plan) will introduce variable delay so we still need some maximum value (of temporal mis-alignment between reference and processed videos) that we will consider in this test plan (See Section 7.2.4)
- The proposed replacement text does not address this issue. It seems to focus only on constant delay.

About the re-scaling issue:

- It was agreed that vertical and horizontal re-scaling is not allowed.
- We propose to keep this decision, mainly because image re-scaling is not really linked to image quality.
- Take the example of a tennis game shot in 4:3 aspect ratio and then displayed on a 16:9 screen. One has the choice to display the image with the correct aspect ratio (4:3) with black bars, or to stretch the image to fill the 16:9 display. Some people will hate watching the programme in the wrong aspect ratio, whilst others will prefer to fill up their big 16:9 screen even if players look squashed… This has nothing to do with image quality.

Section 9 – Objective Model Evaluation Criteria

In v.2.2 of the HDTV test plan, a proposal is made to use only RMSE as the metric to evaluate model performance.

This proposal has some benefit. However we would strongly suggest that VQEG considers an approach similar to the one used for the ITU-T SG12 P.OLQA competition where a modified RMSE was selected as the only metric to evaluate the performance of candidate models. This improved RMSE takes into account the confidence interval around MOS (which is not the case with the standard RMSE).

The issue with (standard) MSE is that it does not take into account the confidence interval around a subjective MOS, i.e. a video is always characterized by a MOS within a certain confidence interval. Consider the following example:

- A video has a MOS=3.24 with a CI=0.20 (where CI denotes the 95% level confidence interval). In other words, 95 out of 100 times, the ‘real’ subjective value is somewhere between 3.04 (3.24-0.20) and 3.44 (3.24+0.20).
- Model A predicts MOSp=3.30. With a standard MSE, abs(Error)=0.06 for this data point (3.30-3.24)
- Model B predicts MOSp=3.40. With a standard MSE, abs(Error)=0.16 for this data point (3.40-3.24)

- With the standard MSE, it would mean that Model A is more accurate than Model B for this data point. However this is not really true since the statistical interpretation of the subjective data is that the real subjective value is somewhere between 3.04 and 3.44, so both models were able to predict the subjective MOS within the CI.
- With this example, a modified MSE would say that (modified) Error=0 for both Model A and B because their prediction falls within the CI.
- On the other hand, Model C predicts MOSp=3.50. Then (modified) MSE would mean that (modified) Error=3.50-3.44=0.06 for Model C on this data point.
- Similarly, Model D predicts MOSp=2.90. Then (modified) MSE would mean that (modified) Error=3.04-2.90=0.14 for Model D on this data point.

We propose that this "improved RMSE" be considered by VQEG if RMSE is the only metric to be used for primary analysis. The benefit is that it keeps the discrimination power of a standard RMSE but also takes into account the inherent variability around a subjective score. This makes it a fairer evaluation criterion for objective models.

Section 9 – Objective Model Evaluation Criteria

In v.2.2 of the HDTV test plan, a proposal is made to use DMOS for evaluation of all model types.

Using DMOS for FR/RR models and MOS for NR models does complicate the data analysis but it does not make sense to use DMOS to evaluate NR models.

We propose to keep decision to use DMOS for FR/RR models and MOS for NR models.

Section 9.4 – Mapping to the Subjective Scale

We support the option using the cubic monotonic polynomial fit.

Section 9.6 - Averaging Process

If a test does not follow a rectangular design (a rectangular design refers to a design where all SRCS are processed through all HRCs), then secondary analysis per HRC (by averaging all SRCs associated with the same HRC) should not be allowed for that test. See previous comment regarding Section 4.3 (Test Design).